Análisis psicométrico de exámenes de Matemáticas y Lenguaje y Comunicación en CONALEP Estado de México

Psychometric analysis of Mathematics and Language and Communication exams in CONALEP State of Mexico.

Mtro. Irvin Rodolfo Tapia Bernabé[1]
Lic. Luis Enrique González Mejía[2]

Resumen. El presente trabajo demuestra en qué medida los reactivos utilizados en pruebas censales de tipo diagnóstico de Matemáticas y en Lenguaje y Comunicación, impacta el desempeño de estudiantes de segundo semestre de CONALEP Estado de México. Como objetivo específico, se tuvo medir el grado en el que los ítems son capaces de establecer diferencias en estudiantes con niveles altos o bajos en sus habilidades, capacidades y conocimientos, así como clasificar por nivel de dificultad cada uno de los ítems de los exámenes de diagnóstico, ambos bajo la perspectiva de la Teoría Clásica del Test. La investigación se encuentra centrada bajo un enfoque cuantitativo con un diseño descriptivo transeccional. Se llegó a la conclusión de que los índices altos de dificultad y bajos de discriminación de los reactivos de ambos exámenes afectaron el desempeño de los estudiantes en su evaluación diagnóstica.

Palabras clave: Análisis psicométrico, pruebas censales, dificultad y discriminación.

1 Colegio de Educación Profesional Técnica del Estado de México. Jefe de Proyecto de Evaluación Educativa irtb.tapia@gmail.com

2 Colegio de Educación Profesional Técnica del Estado de México. Jefe de Proyecto de Evaluación Educativa. glezenkike@gmail.com

Abstract. The present research work shows the extent to which the reagents used in census tests of the diagnostic type of Mathematics and in Language and Communication, impacted the performance of second-year students of CONALEP Estado de México. As a specific objective, we had to measure the degree to which the items are able to establish differences in students with high or low levels in their skills, abilities and knowledge, as well as to classify by level of difficulty each one of the items of the exams. diagnosis, both from the perspective of the Classical Test Theory. The research is centered on a quantitative approach with a descriptive transectional design. It was concluded that the high levels of difficulty and low discrimination of the reagents of both exams affected the performance of the students in their diagnostic evaluation.

Key Words: Psychometric analysis, census tests, difficulty and discrimination.

Planteamiento del problema

Objetivos

Objetivo General:

Analizar los reactivos de la prueba censal de Matemáticas y de Lenguaje y Comunicación de segundo semestre en CONALEP Estado de México, a través de las teorías psicométricas a fin de determinar la calidad de los instrumentos utilizados.

Objetivos específicos:

- Establecer los valores equivalentes a 0 y 1, obtenidos a partir de los resultados de la calificación de los reactivos de opción múltiple.
- Determinar los resultados de los indicadores de calidad para pruebas de opción múltiple que componen la prueba de conocimientos de matemáticas y lenguaje y comunicación.
- Determinar el grado en que se relacionan los reactivos de la prueba, para conocer la consistencia interna del instrumento empleado.

Pregunta de investigación

¿Cuál es el modelo psicométrico apropiado para determinar los indicadores de calidad en exámenes de opción múltiple?

¿De qué manera impactó la calidad de los reactivos de los exámenes en la evaluación de los estudiantes?

Justificación de la investigación

Elevar la calidad de la educación en México, es hoy en día uno de los principales ejes de atención de la política pública en nuestro país. A raíz de la creación de la Política Nacional de Evaluación de la Educación (INEE, 2015) se pretendió entre otros, desarrollar una cultura de evaluación y toma de decisiones basada en la evidencia, así como la construcción de los Programas Estatales de Evaluación y Mejora Educativa (PEEME).

Las áreas de evaluación en los estados se encuentran en proceso de fortalecimiento de sus capacidades institucionales y han desarrollado, de la mano del Instituto Nacional para la Evaluación de la Educación (INEE), sus propios Programas Estatales de Evaluación y Mejora Educativa (PEEME).

En el Estado de México, se ha integrado al PEEME de la EMS, el desarrollo de propuestas de evaluación para la mejora de los procesos a través del desarrollo de exámenes realizados por sus cuerpos académicos (CESPD, s.f.). Un ejemplo de lo anterior es el instrumento para evaluar las competencias Matemáticas y Lenguaje y Comunicación en estudiantes de segundo semestre de CONALEP Estado de México. Si bien este tipo hechos representa acciones plausibles ante este panorama, es importante cuestionarse: ¿Cumple este tipo de instrumentos de evaluación los estándares de calidad establecidos por las organizaciones internacionales? así como ¿De qué manera impactó la calidad de los reactivos de los exámenes en la evaluación de los estudiantes?

Cuando se utilizan instrumentos de gran escala y alto impacto los cuales son diseñados para aplicarse en más de un plantel escolar, como en el caso de las pruebas censales, es necesario conocer los indicadores técnicos que definen la calidad del instrumento educativo. Por su

dimensión y por el poderoso impacto social que tienen, su elaboración debe ajustarse a rigurosos estándares de calidad (Aiken, 1996).

El propósito de una prueba educativa es hacer inferencias acerca del conocimiento que tiene un estudiante respecto al dominio evaluativo que se pretende medir; información que es útil a los educadores para tomar decisiones tendientes a mejorar el proceso educativo (INEE, 2005). Sin embargo, sin importar qué tan cuidadosamente se elabore una prueba, los resultados no tienen ningún valor si esta no se administra y califica en forma adecuada. (Lewis R., 2013).

Una fuente importante de estos recursos son los Estándares para la Evaluación Educativa y Psicológica, una serie de 264 normas para construir, evaluar, administrar, calificar, interpretar y usar los resultados. Enfatizan la importancia de tomar en cuenta el bienestar de las personas que hacen una prueba y evitar el mal uso de los instrumentos de evaluación (Backoff, Larrazolo, & Rosas, 2000). Sin embargo, mientras que, en países desarrollados, es obligatorio que estos criterios de calidad se satisfagan, en México es inexistente esta normatividad.

El presente trabajo de investigación educativa busca examinar los resultados psicométricos relacionados con el nivel de dificultad y poder de discriminación de los reactivos de la evaluación diagnóstica dirigida a estudiantes de segundo semestre de la generación 2017-2020 en el Conalep Estado de México, guiado mediante la Teoría Clásica de los Test, vigente de acuerdo con (Backoff, Larrazolo, & Rosas, 2000).

El estudio generará precedentes en la valoración de la calidad de los instrumentos de medición empleados para evaluar el desarrollo de las competencias en estudiantes, a fin de proporcionar información correcta para la toma de decisiones. Además, aporta una guía a la labor de evaluación educativa de las instituciones de Educación Media Superior a través de la experiencia de Conalep Estado de México.

Fundamentación teórica

Introducción

Dado que el presente trabajo se encuentra centrado en el análisis de los reactivos de opción múltiple utilizados en pruebas de evalua-

ción de los aprendizajes en Matemáticas y Lenguaje y Comunicación, es fundamental dar a conocer los ejes conceptuales para apoyar a la lectura del lenguaje técnico, contenido en el cuerpo metodológico del presente trabajo.

En este sentido, comenzaremos por definir que el análisis de reactivos según (INEE, 2013) forma parte del proceso de validación de un instrumento en el desarrollo de exámenes a través de pruebas objetivas estandarizadas, en donde su ejecución principalmente nos permitirá:

1. Mejorar la calidad de los reactivos, ya que el análisis no solo aporta información de los alumnos, si no específicamente de los reactivos.
2. Aporta información útil para retroalimentar puntos difíciles y errores generalizados por grupos de alumnos.
3. Ayuda a establecer criterios de evaluación, despreciando reactivos de mala calidad.
4. Generar bancos de reactivos de calidad para su uso posterior.

Para lograr la construcción de exámenes de evaluación adecuados, la Teoría Clásica de los Test (TCT) en las últimas décadas ha sido una guía para los constructores de pruebas. Esta teoría considera que, tras aplicar y calificar una prueba, se obtiene el puntaje que logró cada examinado, el cual es conocido como puntuación observada.

Sin embargo, el nivel real de dominio o habilidad que tiene el estudiante no lo conocemos; porque la puntuación que observamos está influenciada por diversos errores de medición. La TCT nos ayuda a reducir los errores que cometemos al construir un examen, de modo que sea posible conocer el nivel de habilidad real de los estudiantes.

Teoría clásica de los test

La TCT ha sido el modelo dominante en la teoría de test, y tiene aún una vigencia representativa en el campo de la evaluación psicológica y educativa. Esta teoría propuesta por Charles Sperman a inicios del siglo XX, usa un modelo matemático sustentado en la curva normal que supone que la habilidad de un sujeto es la sumatoria de los puntajes obtenidos al responder una serie de ítems de una prueba.

La TCT se centra en la estimación del puntaje de una persona como si esta hubiera respondido al universo total de preguntas posibles, como este universo es infinito, es necesario hacer una estimación de este puntaje, el cual tendrá cierta cantidad de error (Navas, 1994).

Lo anterior de acuerdo con (Muñiz, 2010) se define los siguientes supuestos en los que esta soportada la TCT:

1. El puntaje verdadero de una persona en un test es igual a aquella puntuación que obtendría como media si se le pasase infinitas veces el test.
2. El valor de la puntuación verdadera de una persona no tiene relación con el error que afecta esa puntuación, es decir, puede haber puntuaciones verdaderas altas con errores bajos, o altos, no hay conexión entre el tamaño de la puntuación verdadera y el tamaño de los errores.
3. Los errores de medida de las personas en un test no están relacionados con los errores de medida en otro test distinto.

Índice de dificultad

El índice de dificultad de un ítem de acuerdo con la TCT, se entiende como la proporción de personas que responden correctamente un reactivo de una prueba. Entre mayor sea esta proporción, menor será su dificultad. Lo que quiere decir que se trata de una relación inversa: a mayor dificultad del ítem, menor será (Backoff, Larrazolo, & Rosas, 2000).

Para calcular la dificultad de un ítem, se divide el número de personas que contestan correctamente el ítem entre el número total de personas que intentan resolver el ítem (correcta o incorrectamente).

$$ID = \frac{Ai}{Ni} \text{ .....(Ecuación 1)}$$

Donde:

A_i = Número de personas que aciertan el ítem.

N_i = Número de personas que intentaron resolver el reactivo.

Para la evaluación de los reactivos se emplea la tabla 01.

Tabla 01. Interpretación del nivel de dificultad

Grado de dificultad del ítem	Interpretación
85	Muy difícil
60-85	Difícil
40-60	Moderado
15-40	Fácil
15	Muy Fácil

Índice de discriminación

Si la prueba y un ítem miden la misma habilidad o competencia, podemos esperar que quien tuvo una puntuación alta en todo el test deberá tener altas probabilidades de contestar correctamente el ítem. También debemos esperar lo contrario, es decir, que quien tuvo bajas puntuaciones en el test, deberá tener pocas probabilidades de contestar correctamente el reactivo. Así, un buen ítem debe discriminar entre aquellos que obtuvieron buenas calificaciones en la prueba y aquellos que obtuvieron bajas calificaciones.

Para determinar el índice discriminativo de un ítem, utilizaremos la siguiente fórmula:

$$P = \frac{Ac - Ai}{M} \ldots\ldots \text{(Ecuación 2)}$$

Donde: Ac: La frecuencia de aciertos en el grupo superior (convenientes)

Ai: La frecuencia de aciertos en el grupo inferior (inconvenientes)

M: El total de individuos en cada grupo

Entre más alto es el índice de discriminación, el reactivo diferenciará mejor a las personas con altas y bajas calificaciones. Si todas las personas del Ac contestan correctamente un reactivo y todas las personas del Ai contestan incorrectamente, entonces P = 1 (valor máximo de este indicador); si sucede lo contrario, P = -1 (valor máximo negativo); si ambos grupos contestan por igual, P = 0 (valor mínimo de discriminación).

Ebel y Frisbie como se cita en (Backoff, Larrazolo, & Rosas, 2000) nos dan la siguiente regla de "dedo" para determinar la calidad de los reactivos, en términos del índice de discriminación. La Tabla 1, muestra los valores D y su correspondiente interpretación. Asimismo, en la tabla 2 se señalan las recomendaciones para cada uno de estos valores.

Tabla 02. Indice de discriminación de los reactivos según su valor

D=	Calidad	Recomendaciones
>0,39	Excelente	Conservar
0,30 . 0,39	Buena	Posibilidades de mejorar
0,20 . 0,29	Regular	Necesidad de revisar
0,00 . 0,20	Pobre	Descartar o revisar a profundidad
<-0,01	Pésima	Descartar definitivamente

Coeficiente de correlación biserial

El coeficiente de correlación biserial se calcula para determinar el grado en que las competencias que mide el test también las mide el reactivo. Proporciona una estimación de la correlación producto-momento de Pearson entre la calificación total de la prueba y el continuo hipotético del reactivo, cuando éste se dicotomiza en respuestas correctas e incorrectas.

Se consideran aceptables los ítems con valores superiores a 0.25. La ecuación para obtener este indicador, es la siguiente:

$$r_{bp} = \frac{\bar{x}_p - \bar{x}}{\sigma} * \sqrt{\frac{p}{q}} \quad \text{..... (Ecuación 3)}$$

Donde:

p: La proporción de individuos que acertaron

q: La proporción de individuos que fallaron

Xp: La media en X de los sujetos cuya proporción es *p*

X: La media del test

Sx: La desviación típica del test

Interpretación de los resultados empíricos de la Teoría Clásic del Test

De acuerdo con (INEE, 2013) los resultados obtenidos en el análisis empírico de los indicadores de calidad según la TCT, nos permitirá conocer las dificultades que tienen frecuentemente los estudiantes al responder un reactivo, las cuales pasan inadvertidas al especialista que lo construyó. Entre los problemas más comunes se encuentran los siguientes:

1. La base del reactivo es confusa debido al uso de lenguaje sofisticado, estructura gramatical compleja, o a su mala formulación.
2. La pregunta resulta muy fácil debido a que los distractores son muy obvios, poco plausibles o mal redactados.
3. a pregunta es muy difícil debido a la "cercanía conceptual" de los distractores, lo cual hace difícil discriminar la respuesta correcta o que haya más de una respuesta correcta, etcétera.

Por lo anterior, es claro que la única forma de conocer si el reactivo de un examen influye en el desempeño de un estudiante, es a través del análisis empírico de sus resultados. Los resultados generalmente aportan importante para la mejora de la calidad de los instrumentos, asegura que los reactivos estén formulados correctamente y cumplan con su propósito en la prueba correspondiente.

Metodología

Paradigma de la investigación

Se encuentra basado en la corriente positivista, tiene un fundamento empirico que se desarrolla a través del uso de fuentes estadísticas para el análisis de la validez, fiablidad y objetividad de los instrumentos empleados en la evaluación diagnóstica.

Enfoque de la metodología

Contempla un enfoque cuantitativo, a partir del método hipotético-deductivo, emplea una metodología estructurada y formal para la recolección de datos, la medición numérica y el análisis estadístico para establecer pautas de comportamiento a través de los supuestos de la Teoría Clásica del Test.

Método de la investigación

De acuerdo (Hernández, Fernández, & Baptista, 2014) el estudio se caracteriza por tener un alcance descriptivo y un diseño no experimental del tipo transeccional descriptivo debido a que las variables no serán manipuladas. Se fundamenta en que solo se emplearán los datos recolectados posterior a la aplicación de dos pruebas de conocimientos a estudiantes de una generación escolar.

Técnica de recolección de datos

A partir de los exámenes desarrollados por el CONALEP Estado de México, los cuales se denominaron: Examen de Evaluación Diagnóstica en Matemáticas y Examen de Evaluación Diagnóstica en Lenguaje y Comunicación. Ambos tuvieron como objetivo valorar el dominio de contenidos adquiridos en el primer semestre (ver anexo A).

Las pruebas contienen reactivos de tipo politómico o mejor conocidos como opción múltiple, los cuales se caracterizan por su versatilidad para evaluar conocimiento factual (puramente memorístico), habilidades intelectuales de alto orden, o disposiciones actitudinales y valorativas. Con ese tipo de preguntas, siempre que sean bien utilizadas, se puede medir una gran cantidad de atributos sofisticados de los estudiantes (Instituto Nacional de Evaluación Educativa, 2013).

Los instrumentos incluyen los siguientes aspectos evaluados:

Matemáticas: "Sentido numérico y pensamiento algebraico" con 13 reactivos, "Cambios y relaciones", con 2 reactivos. Conformando un total de 15 reactivos.

Lenguaje y Comunicación:"Texto expositivo" con 11 reactivos, "Manejo y construcción de la información" 3 reactivos, "Texto literario" con un reactivo.

Selección de la muestra

Se consideró a la totalidad de los resultados obtenidos en la aplicación censal de los exámenes de diagnóstico en las competencias en Matemáticas y Lenguaje y comunicación en alumnos de segundo semestre de la generación 2017-2020 de los 39 planteles del Conalep, como a continuación se indica:

- Examen de Matemáticas: 14,392 estudiantes
- Examen de Lenguaje y Comunicación: 13,773 estudiantes

Procedimiento

Los pasos para la administración y calificación del examen fueron definidos por el área académica del CONALEP Estado de México como a continuación se indica: (1) se compartió a planteles el link el examen el cual fue aplicado a través de un formulario de Google Drive; (2) los responsables de la aplicación del examen en planteles colocaron las pantallas de inicio a la prueba en los laboratorios de informática; (3) los estudiantes respondieron el examen sin ningún tipo de ayuda (calculadora, diccionario libros, etc.). Durante este proceso, se encontró siempre presente una persona capacitada que resolvió cualquier problema o duda sobre el manejo de la parte computarizada del examen.

Las respuestas de los estudiantes fueron extraídas de la aplicación y conjuntadas en una base de datos para su procesamiento estadístico. Se generó una hoja de cálculo en Excel para su interpretación de las respuestas a un formato binario (0 y 1). Utilizando el software de hoja de cálculo Excel, se calcularon los valores de p (dificultad), D (índice de discriminación) y rpbis para todos los reactivos del examen.

El índice de dificultad se calculó con la (Ecuación 1), el índice de discriminación con la (Ecuación 2) y el coeficiente de discriminación con la (Ecuación 3).

	Matemáticas			Lenguaje y Comunicación		
Ítem	Índice de Dificultad	Índice de Discriminación	Coeficiente de Correlación Biserial	Índice de Dificultad	Índice de Discriminación	Coeficiente de Correlación Biserial
1	60.7	0.48	0.40	40.10	0.44	0.36
2	68.7	0.37	0.33	56.93	0.62	0.50
3	72.5	0.42	0.33	37.87	0.47	0.38
4	80.5	0.29	0.33	84.56	0.14	0.17
5	71.7	0.31	0.30	75.92	0.33	0.34
6	61.4	0.36	0.31	78.07	0.20	0.21
7	71.0	0.35	0.33	74.60	0.35	0.34
8	48.0	0.51	0.40	67.55	0.48	0.41
9	81.3	0.27	0.33	72.29	0.17	0.15
10	82.7	0.22	0.28	46.05	0.50	0.40
11	69.9	0.35	0.30	70.56	0.32	0.31
12	84.3	0.19	0.24	63.15	0.34	0.30
13	77.0	0.30	0.31	79.64	0.17	0.17
14	55.9	0.39	0.29	92.16	0.05	0.10
15	70.9	0.21	0.20	81.48	0.13	0.15

Resultados

La tabla 03 muestra el índice de dificultad, índice de discriminación y coeficiente de correlación biserial de los 15 reactivos de ambas pruebas.

Tabla 03. índice de dificultad, índice de discriminación y coeficiente de correlación biserial

	Lenguaje y Comunicación	Matemáticas
Ítem	Interpretación	Interpretación
Ítem 1	Moderado	Difícil
Ítem 2	Moderado	Difícil
Ítem 3	Fácil	Difícil
Ítem 4	Difícil	Difícil
Ítem 5	Difícil	Difícil
Ítem 6	Difícil	Difícil
Ítem 7	Difícil	Difícil
Ítem 8	Difícil	Moderado
Ítem 9	Difícil	Difícil
Ítem 10	Moderado	Difícil
Ítem 11	Difícil	Difícil
Ítem 12	Difícil	Difícil
Ítem 13	Difícil	Difícil
Ítem 14	Muy dificil	Moderado
Ítem 15	Difícil	Difícil

Índice de dificultad: Los resultados arrojados del examen de Matemáticas el 86%, de los reactivos se ubican en un nivel de dificultad

alto y 14% moderado. En el caso de Lenguaje y Comunicación, indican que el 73% de los reactivos tienen un índice de dificultad alto, el 20% moderado y un 10% fácil. La tabla 04 muestra la interpretación de la dificultad de los ítems.

Tabla 04. Interpretación de la dificultad de los ítems.

	Lenguaje y Comunicación		Matemáticas	
Ítem	Calidad	Recomendación	Calidad	Recomendación
Ítem 1	Excelente	Conservar	Excelente	Conservar
Ítem 2	Excelente	Conservar	Buena	Posibilidades de mejorar
Ítem 3	Excelente	Conservar	Excelente	Conservar
Ítem 4	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar
Ítem 5	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 6	Pobre	Descartar o revisar a profundidad	Buena	Posibilidades de mejorar
Ítem 7	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 8	Excelente	Conservar	Excelente	Conservar
Ítem 9	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar
Ítem 10	Excelente	Conservar	Regular	Necesidad de revisar
Ítem 11	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 12	Buena	Posibilidad de mejorar	Pobre	Descartar o revisar a profundidad
Ítem 13	Pobre	Descartar o revisar a profundidad	Buena	Posibilidad de mejorar
Ítem 14	Pobre	Descartar o revisar a profundidad	Buena	Posibilidad de mejorar
Ítem 15	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar

Índice de discriminación: En el caso de Matemáticas el 20% de los reactivos presenta una excelente calidad, el 46% en una calidad buena, el 26.6 % una calidad regular y el 8% una mala calidad. En el examen de Lenguaje y Comunicación se identificó el 33% de los reactivos con una excelente calidad, el 26.6% con una buena calidad, el 4.8% con calidad regular y 26.6 con mala calidad. La tabla 05 muestra la evaluación de la calidad de los ítems.

Tabla 05. Calidad de ítems

	Lenguaje y Comunicación		Matemáticas	
Ítem	Calidad	Recomendación	Calidad	Recomendación
Ítem 1	Excelente	Conservar	Excelente	Conservar
Ítem 2	Excelente	Conservar	Buena	Posibilidades de mejorar
Ítem 3	Excelente	Conservar	Excelente	Conservar
Ítem 4	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar
Ítem 5	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 6	Pobre	Descartar o revisar a profundidad	Buena	Posibilidades de mejorar
Ítem 7	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 8	Excelente	Conservar	Excelente	Conservar
Ítem 9	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar
Ítem 10	Excelente	Conservar	Regular	Necesidad de revisar
Ítem 11	Buena	Posibilidad de mejorar	Buena	Posibilidades de mejorar
Ítem 12	Buena	Posibilidad de mejorar	Pobre	Descartar o revisar a profundidad
Ítem 13	Pobre	Descartar o revisar a profundidad	Buena	Posibilidad de mejorar
Ítem 14	Pobre	Descartar o revisar a profundidad	Buena	Posibilidad de mejorar
Ítem 15	Pobre	Descartar o revisar a profundidad	Regular	Necesidad de revisar

Coeficiente de correlación biserial: En Matemáticas el 86.6% muestra una correlación por encima de 0.25 dentro de la clasificación aceptable. Para el caso de Lenguaje y Comunicación el 60% de los reactivos se situaron por encima del 0.25 de la correlación.

Valoración de ítems

En la valoración de ítems se describen cómo funcionó una pregunta en una situación dada, es decir; no necesariamente se asocia a juicios de valor sobre la calidad de la pregunta al valor de estos índices. A continuación, se refleja las inferencias para cada reactivo a partir de la estadística descriptiva de cada instrumento:

Examen de matemáticas:

Ítem 1: El 60.7% contestaron mal al reactivo y el 39.30% lo hizo correctamente. Presenta un índice de discriminación de 0.48 y un coeficiente de correlación biserial de 0.40. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Excelente" lo cual indica poder ser conservado.

Ítem 2: El 68.7% respondieron mal al reactivo y el 31.3% lo hizo correctamente. Su índice de discriminación es de 0.37 y un coeficiente de correlación biserial del 0.33. El ítem presenta un nivel "Difícil" y con calidad "Buena", lo cual indica la posibilidad de mejorar.

Ítem 3: El 72.5% respondieron mal al reactivo, el 27.5% lo hizo correctamente. Su índice de discriminación es de 0.42 y un coeficiente de correlación biserial de 0.33. El ítem presenta un nivel "Difícil" y tiene una calidad "Buena", lo cual indica poder ser conservado.

Ítem 4: El 80.5% respondieron mal al reactivo, el 19.5% lo hizo correctamente. Su índice de discriminación es de .29 y tiene un coeficiente de correlación biserial de .33. El ítem presenta un nivel "Dif**ícil" y tiene una calidad "Regular**", lo cual indica la necesidad de revisar.

Ítem 5: El 71.7% respondieron mal al reactivo, el 28.3% lo hizo correctamente. Su índice de discriminación es de 0.31 y tiene un coeficiente de correlación biserial de 0.30. El ítem presenta un

nivel “Difícil” y tiene una calidad “Buena”, lo cual indica la posibilidad de mejorar.

Ítem 6: El 61% respondieron mal al reactivo, el 39% lo hizo correctamente. Su índice de discriminaci**ón es de** 0.36 y tiene un coeficiente de correlación biserial de 0.31. El ítem presenta un nivel “Difícil” y tiene una calidad “Buena”, lo cual indica la posibilidad de mejorar.

Ítem 7: El 71.4% respondieron mal al reactivo, el 38.6% lo hizo correctamente. Su índice de discriminaci**ón es de** 0.35 y tiene un coeficiente de correlación biserial de 0.33. El ítem presenta un nivel “Difícil” y tiene una calidad “Buena”, lo cual indica la posibilidad de mejorar.

Ítem 8: El 48% respondieron mal al reactivo, el 52% lo hizo correctamente. Su índice de discriminación es de 0.51 y tiene un coeficiente de correlación biserial de 0.40. El ítem presenta un nivel de dificultad “Moderado” y tiene una calidad “Excelente”, lo cual indica poder ser conservado.

Ítem 9: El 81.3% respondieron mal al reactivo, el 18.7% lo hizo correctamente. Su índice de discriminación es de 0.27 y tiene un coeficiente de correlación biserial de 0.33. El ítem presenta un nivel de “Difícil” y tiene una calidad “Regular”, lo cual indica la necesidad de revisar.

Ítem 10: El 82.7% respondieron mal al reactivo, el 17.3% lo hizo correctamente. Su índice de discriminación es de 0.22 y tiene un coeficiente de correlación biserial de 0.28. El ítem presenta un nivel de “Difícil” y tiene una calidad “regular”, lo cual indica la necesidad de revisar.

Ítem 11: El 69.9% respondieron mal al reactivo, el 30.1% lo hizo correctamente. Su índice de discriminación es de 0.35 y tiene un coeficiente de correlación biserial de 0.30. El ítem presenta un nivel de “Difícil” y tiene una calidad “Buena”, lo cual indica la posibilidad de mejorar.

Ítem 12: El 84.3% respondieron mal al reactivo, el 15.7% lo hizo correctamente. Su índice de discriminación es de 0.19 y tiene

un coeficiente de correlación biserial de 0.24. El ítem presenta un nivel de "Difícil" y tiene una calidad "Pobre", lo cual indica descartar o revisar a profundidad.

Ítem 13: El 77% respondieron mal al reactivo, el 23% lo hizo correctamente. Su índice de discriminación es de 0.30 y tiene un coeficiente de correlación biserial de 0.31. El ítem presenta un nivel de "Difícil" y tiene una calidad "Buena", lo cual indica la posibilidad de mejorar.

Ítem 14: El 55.9% respondieron mal al reactivo, el 44.1% lo hizo correctamente. Su índice de discriminación es de 0.39 y tiene un coeficiente de correlación biserial de 0.29. El ítem presenta un nivel de "Moderado" y tiene una calidad "Buena", lo cual indica la posibilidad de mejorar.

Ítem 15: El 70.9% respondieron mal al reactivo, el 29.1% lo hizo correctamente. Su índice de discriminación es de 0.21 y tiene un coeficiente de correlación biserial de 0.20. El ítem presenta un nivel de "Difícil" y tiene una calidad "Buena", lo cual indica la necesidad de revisar.

Examen de Lenguaje y Comunicación:

Ítem 1: El 40.1% contestaron mal al reactivo y el 59.9% lo hizo correctamente. Presenta un índice de discriminación de 0.44 y un coeficiente de correlación biserial de 0.36. Lo anterior coloca al ítem en un nivel de dificultad "Moderado" y con una calidad "Excelente" lo cual indica poder ser conservado.

Ítem 2: El 56.9% contestaron mal al reactivo y el 43.1% lo hizo correctamente. Presenta un índice de discriminación de 0.62 y un coeficiente de correlación biserial de 0.60. Lo anterior coloca al ítem en un nivel de dificultad "Moderado" y con una calidad "Excelente" lo cual indica poder ser conservado.

Ítem 3: El 37.8% contestaron mal al reactivo y el 62.2% lo hizo correctamente. Presenta un índice de discriminación de 0.47 y un coeficiente de correlación biserial de 0.38. Lo anterior coloca al ítem en un nivel de dificultad "Facil" y con una calidad "Excelente" lo cual indica poder ser conservado.

Ítem 4: El 84.5% contestaron mal al reactivo y el 15.5% lo hizo correctamente. Presenta un índice de discriminación de 0.14 y un coeficiente de correlación biserial de 0.17. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Pobre" lo cual indica descartar o revisar a profundidad.

Ítem 5: El 75.9% contestaron mal al reactivo y el 15.5% lo hizo correctamente. Presenta un índice de discriminación de 0.33 y un coeficiente de correlación biserial de 0.34. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Buena" lo cual indica la posibilidad de mejorar.

Ítem 6: El 78% contestaron mal al reactivo y el 22% lo hizo correctamente. Presenta un índice de discriminación de 0.20 y un coeficiente de correlación biserial de 0.21. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Pobre" lo cual indica descartar o revisar a profundidad.

Ítem 7: El 75.9% contestaron mal al reactivo y el 15.5% lo hizo correctamente. Presenta un índice de discriminación de 0.33 y un coeficiente de correlación biserial de 0.34. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Buena" lo cual indica la posibilidad de mejorar.

Ítem 8: El 67.5% contestaron mal al reactivo y el 32.5% lo hizo correctamente. Presenta un índice de discriminación de 0.48 y un coeficiente de correlación biserial de 0.41. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Excelente" lo cual indica poder conservarlo.

Ítem 9: El 72.2% contestaron mal al reactivo y el 27.8% lo hizo correctamente. Presenta un índice de discriminación de 0.17 y un coeficiente de correlación biserial de 0.15. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Pobre" lo cual indica descartar o revisar a profundidad.

Ítem 10: El 46% contestaron mal al reactivo y el 64% lo hizo correctamente. Presenta un índice de discriminación de 0.50 y un coeficiente de correlación biserial de 0.40. Lo anterior coloca al ítem

en un nivel "Moderado" y con una calidad "Excelente" lo cual indica poder ser conservado.

Ítem 11: El 70.5% contestaron mal al reactivo y el 29.5% lo hizo correctamente. Presenta un índice de discriminación de 0.32 y un coeficiente de correlación biserial de 0.31. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Buena" lo cual indica la posibilidad de mejorar.

Ítem 12: El 63.1% contestaron mal al reactivo y el 36.9% lo hizo correctamente. Presenta un índice de discriminación de 0.34 y un coeficiente de correlación biserial de 0.30. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Buena" lo cual indica la posibilidad de mejorar.

Ítem 13: El 79.6% contestaron mal al reactivo y el 20.4% lo hizo correctamente. Presenta un índice de discriminación de 0.17 y un coeficiente de correlación biserial de 0.17. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Pobre" lo cual indica descartar o revisar a profundidad.

Ítem 14: El 92.1% contestaron mal al reactivo y el 7.9% lo hizo correctamente. Presenta un índice de discriminación de 0.05 y un coeficiente de correlación biserial de 0.10. Lo anterior coloca al ítem en un nivel "Muy Difícil" y con una calidad "Pobre" lo cual indica descartar o revisar a profundidad.

Ítem 15: El 81.4% contestaron mal al reactivo y el 18.6% lo hizo correctamente. Presenta un índice de discriminación de 0.13 y un coeficiente de correlación biserial de 0.15. Lo anterior coloca al ítem en un nivel "Difícil" y con una calidad "Pobre" lo cual indica descartar o revisar a profundidad.

En síntesis, el análisis realizado bajo la teoría de la TCT, muestra que ambos exámenes cuentan con un índice de dificultad por arriba del 60%, para efectos de evaluación, los ítems son en promedio "Muy Difíciles" para los estudiantes. En el caso del índice de discriminación, el examen de matemáticas solo presenta un reactivo el cual requiere ser cambiado, sin embargo, en Lenguaje y Comunicación el mismo caso se presenta en 5 de los reactivos. Los resultados de la correlación bise-

rial nos indican en el caso de matemáticas que solamente dos reactivos se encuentran fuera del rango de aceptación, es decir que estos ítems no miden conocimientos de la misma área disciplinar que la mayoría del examen. En Lenguaje y comunicación este coeficiente demuestra que 6 reactivos se encuentran fuera del rango establecido.

Conclusiones

La mejor manera de conocer la calidad de los reactivos de un examen es a través de su análisis empírico una vez puestos a prueba. De acuerdo con la Teoría Clásica del Test aplicada a la evaluación educativa, los indicadores fundamentales para realizar el análisis son el índice de dificultad, el índice de discriminación y la correlación biserial. Esta teoría sigue vigente gracias a la sencillez de sus supuestos matemáticos.

Los supuestos de la Teoría Clásica del Test, aplicado al análisis de reactivos de opción múltiple permite identificar: a) Confusiones que tuvieron los examinados en la elección de su respuesta, debido a la mala redacción del reactivo, al uso de lenguaje sofisticado, estructura gramatical compleja, o a su mala formulación; b) preguntas muy fáciles para los examinados, debido a que los distractores son muy obvios, poco plausibles o mal redactados; c) Preguntas muy difíciles debido a la "cercanía conceptual" de los distractores, lo cual hace difícil discriminar la respuesta correcta.

Los resultados obtenidos en el análisis de la calidad de los reactivos del examen realizado en CONALEP Estado de México, con fines de evaluación diagnóstica a estudiantes de segundo semestre, fue en promedio difícil para la mayoría de los estudiantes. La evidencia empírica obtenida arrojó que, en ambos exámenes, el promedio de acierto fue igual al 30%.

El uso de la TCT permitió identificar omisiones metodológicas en la elaboración del examen, específicamente en la distribución de los niveles de dificultad de los reactivos, así como en la inclusión de reactivos que midieron conocimientos fuera de los objetivos de aprendizaje de los programas de estudio. Lo anterior, de acuerdo con la interpretación de los resultados obtenidos a partir de los supuestos de dicha teoría.

Por tanto, el presente trabajo demuestra la importancia del uso adecuado de las metodologías para la elaboración de exámenes censales por parte de las áreas dedicadas a la evaluación educativa en el sistema educativo de nuestro país. La falta de normatividad al respecto, como se ha demostrado en el presente estudio, conlleva a la generación de información inadecuada para la correcta toma de decisiones.

Así mismo, debe considerarse las necesidades de formación de las áreas de evaluación educativa para hacer frente a las necesidades del Plan Nacional de Evaluación Educativa y el Programa de Evaluación y Mejora Educativa del Estado de México, en razón de que los resultados de las evaluaciones de los aprendizajes deben coadyuvar a la mejora de la calidad educativa.

Finalmente, el presente estudio aporta un referente metodológico para la generación de exámenes de calidad a partir del análisis de reactivos, en virtud de la ausencia de la evaluación para el logro de los aprendizajes (PLANEA) en 2018.

Referencias bibliográficas

BACKHOFF, E., Ibarra, M. A., & Rosas, ,. M. (1995). Sistema computarizado de exámenes (SICODEX). *Revista Mexicana de Psicología*, 55-62.

BACKOFF, E., Larrazolo, N., & Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa*, 1.

CESPD. (s.f.). *cespd.edomex.gob.mx*. Obtenido de http://cespd.edomex.gob.mx/eval_proyectos_eb_peeme

HERNANDEZ Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). *Metodología de la Investigación* . Ciudad de México: McGraw-Hill.

INEE. (septiembre de 2005). Obtenido de www.inee.gob.mx.

INEE. (2013). *www.inee.edu.mx*. Obtenido de Manual técnico para la elaboración de reactivos: http://www.inee.edu.mx/images/stories/Publicaciones/Documentos_tecnicos/De_pruebasymedicion/construccion_reactivos/Partes/construccion06.pdf

INEE. (2015). *Política Nacional de Evaluación Educativa*. Ciudad de México: Documentos Rectores INEE.

INEE. (07 de Diciembre de 2017). *Gob.mx*. Obtenido de https://www.gob.mx/cms/uploads/attachment/file/278497/Calendario_de_Evaluaciones_INEE-SEP_2018.pdf

INEE. (diciembre de 2017). *www.inee.edu.mx*. Obtenido de http://publicaciones.inee.edu.mx/buscadorPub/P1/F/105/P1F105.pdf

INEE. (03 de Marzo de 2013). *http://www.inee.edu.mx*. Obtenido de http://www.inee.edu.mx/images/stories/Publicaciones/Documentos_tecnicos/De_pruebasymedicion/construccion_reactivos/Completo/mtconstrecexcalemarca.pdf

LEWIS R., A. (2013). *Test psicológicos y evaluación* . México: Pearson.

MORALES, P. (05 de mayo de 2009). *Análisis de ítems en pruebas objetivas*. Obtenido de Educrea: https://educrea.cl/wp-content/uploads/2014/11/19-nov-analisis-de-items-en-las-pruebas-objetivas.pdf

MUÑIZ, J. (2010). Las teorías de los test: Teoría clásica y Teoría de respuesta al ítem. *Papeles del psicólogo*, 57-66.

NAVAS, J. M. (1994). Teoría Clásica del Test versus Teoría de Respuesta al Ítem. *Psicológica*, 15.

Fecha de recepción. 18 de mayo de 2018.
Fecha de aceptación. 30 de mayo de 2018.

Anexos

Anexo A.

Competencias específicas evaluadas en las pruebas diagnósticas

No.	Matemáticas	Lenguaje y Comunicación
1	Expresa en lenguaje matemático situaciones donde se desconoce un valor o las relaciones de proporcionalidad entre dos variables, y resuelven problemas que implican proporciones entre cantidades (por ejemplo, el cálculo de porcentajes).	Identifican el tema central, el problema planteado y la estructura de un texto expositivo.
2	Comprender y recordar las operaciones entre conjuntos	Identifican el tema central, el problema planteado y la estructura de un texto expositivo
3	Expresan en lenguaje matemático situaciones donde se desconoce un valor o las relaciones de proporcionalidad entre dos variables.	Deducen el significado de las palabras dentro de un texto
4	Realizan operaciones que involucran números enteros y signos de agrupación	Deducen el significado de las palabras dentro de un texto.
5	Realizan operaciones que involucran números enteros y signos de agrupación	Identifican el concepto de referente
6	Resuelven problemas aditivos con fracciones con denominador común	Identifican el concepto de referente
7	Resuelven problemas multiplicativos de fracciones mixtas.	Realizan la lectura crítica de textos expositivos.
8	Realizan operaciones de conjuntos que involucran el planteamiento de problemas y soluciones gráficas.	Planean la estructura de un texto de acuerdo con un propósito comunicativo.
9	Realizan operaciones que involucran números enteros y signos de agrupación	Distinguen las características de distintos tipos de texto.
10	Expresan en lenguaje matemático situaciones donde se desconoce un valor o las relaciones de proporcionalidad entre dos variables	Diferencian los propósitos comunicativos de este tipo texto y otros.
11	Realizan la suma de funciones y evalúan números positivos en ellas.	Distinguen las características de distintos tipos de texto
12	Aplicar los productos notables	Identifican las características de un texto narrativo sencillo
13	Interpretan las relaciones y parámetros de la función lineal dentro una situación.	Identifican el tema central, el problema planteado y la estructura de un texto expositivo.
14	Identifica definiciones de conjuntos en términos simbólicos o algebraicos	Deducen el significado de las palabras dentro de un texto.
15	Resuelven problemas de valor faltante en tablas manejando números reales	Identifican el tema central, el problema planteado y la estructura de un texto expositivo.